Accurate and Explainable Image-Based Prediction Using a Lightweight Generative Model

Mauri, Chiara; Cerri, Stefano; Puonti, Oula; Mühlau, Mark; Van Leemput, Koen

Published in:
Medical Image Computing and Computer Assisted Intervention – MICCAI 2022

Link to article, DOI:
10.1007/978-3-031-16452-1_43

Publication date:
2022

Document Version
Peer reviewed version

Accurate and Explainable Image-based Prediction Using a Lightweight Generative Model

Chiara Mauri[1⋆], Stefano Cerri[2], Oula Puonti[3], Mark Mühlau[4], and Koen Van Leemput[1,2]

[1] Department of Health Technology, Technical University of Denmark, Denmark
[2] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[3] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark
[4] Department of Neurology and TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Germany

Abstract. Recent years have seen a growing interest in methods for predicting a variable of interest, such as a subject's age, from individual brain scans. Although the field has focused strongly on nonlinear discriminative methods using deep learning, here we explore whether linear generative techniques can be used as practical alternatives that are easier to tune, train and interpret. The models we propose consist of (1) a causal forward model expressing the effect of variables of interest on brain morphology, and (2) a latent variable noise model, based on factor analysis, that is quick to learn and invert. In experiments estimating individuals' age and gender from the UK Biobank dataset, we demonstrate competitive prediction performance even when the number of training subjects is in the thousands – the typical scenario in many potential applications. The method is easy to use as it has only a single hyperparameter, and directly estimates interpretable spatial maps of the underlying structural changes that are driving the predictions.

1 Introduction

Image-based prediction methods aim to estimate a variable of interest, such as a subject's diagnosis or prognosis, directly from a medical scan. Predicting a subject's age based on a brain scan – the so called brain age – in particular has seen significant interest in the last decade [12], with the gap between *brain* age and *chronological* age being suggested as a potential biomarker of healthy aging and/or neurological disease [12, 25].

Methods with state-of-the-art prediction performance are currently based on *discriminative learning*, in which a variable of interest x is directly predicted from an input image $\boldsymbol{t}$. Although there are ongoing controversies in the literature regarding whether nonlinear or linear discriminative methods predict better [23,

⋆ Corresponding author. Email address: cmau@dtu.dk

32, 28], recent years have seen a strong focus on nonlinear variants based on deep learning (DL), with impressive performances especially when the training size is very large [28]. Nevertheless, these powerful methods come with a number of potential limitations:

The available training size is often limited: While methods for predicting age and gender can be trained on thousands of subjects using large imaging studies [4, 21, 24, 14, 16], in many potential applications the size of the training set is much more modest. In a recent survey on single-subject prediction of brain disorders in neuroimaging, the mean and median samples size was only 186 and 88 subjects, respectively [5]. Even in such ambitious imaging projects as the UK Biobank [4], the number of subjects with diseases such as multiple sclerosis is only projected to be in the hundreds in the coming years.

Discriminative methods are hard to interpret: As opposed to *generative* methods that explicitly model the effect a variable of interest x has on a subject's image $\boldsymbol{t}$, correctly interpreting the internal workings of discriminative methods is known to be difficult [22, 6, 20, 3]. Whereas the spatial weight maps of linear discriminative methods, or more generally the saliency maps of nonlinear ones [35, 8, 17, 34, 38, 37, 33, 36], are useful for highlighting which image areas are being used in the prediction process [20, 29], they do not explain *why* specific voxels are given specific attention: Amplifying the signal of interest, or suppressing noninteresting noise characteristics in the data [22].

DL can be more difficult to use: Compared to less expressive techniques, DL methods are often harder to use, as they can be time consuming to train, and have many more "knobs" that can be turned to obtain good results (e.g., the choice of architecture, data augmentation, optimizer, training loss, etc. [28]).

In this paper, we propose a lightweight generative model that aims to be easier to use and more straightforward to interpret, without sacrificing prediction performance in typical sample size settings. Like in the mass-univariate techniques that have traditionally been used in human brain mapping [7, 13, 11, 18], the method has a causal forward model that encodes how variables of interest affect brain shape, and is therefore intuitive to interpret. Unlike such techniques, however, the method also includes a linear-Gaussian latent variable noise model that captures the dominant correlations between voxels. As we will show, this allows us to efficiently "invert" the model to obtain accurate predictions of variables of interest, yielding an effective linear prediction method without externally enforced interpretability constraints [9, 39].

The method we propose can be viewed as an extension of prior work demonstrating that naive Bayesian classifiers can empirically outperform more powerful methods when the training size is limited, even though the latter have asymptotically better performance [15, 27]. Here we show that these findings translate to prediction tasks in neuroimaging when the strong conditional independence assumption of such "naive" methods is relaxed. Using experiments on age and gender prediction in the UK Biobank imaging dataset, we demonstrate

empirically that, even when the number of training subjects is the thousands, our lightweight linear generative method yields prediction performance that is competitive with state-of-the-art nonlinear discriminative [28], linear discriminative [31], and nonlinear generative [40] methods.

2 Method

Let $\boldsymbol{t}$ denote a vectorized version of a subject's image, and $\boldsymbol{\phi} = (x, \boldsymbol{\phi}_{\setminus x}^T)^T$ a vector of variables specific to that subject, consisting of a variable of interest x (such as their age or gender), along with any other known[5] subject-specific covariates $\boldsymbol{\phi}_{\setminus x}$. A simple generative model is then of the form

$$\boldsymbol{t} = \boldsymbol{W}\boldsymbol{\phi} + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{\eta}$ is a random noise vector, assumed to be Gaussian distributed with zero mean and covariance $\boldsymbol{C}$, and $\boldsymbol{W} = (\boldsymbol{w}_x \ \boldsymbol{W}_{\setminus x})$ is a matrix with spatial weight maps stacked in its columns. The first column, $\boldsymbol{w}_x$, expresses how strongly the variable of interest x is expressed in the voxels of $\boldsymbol{t}$; we will refer to it as the *generative* weight map. Taking everything together, the image $\boldsymbol{t}$ is effectively modeled as Gaussian distributed:

$$p(\boldsymbol{t}|\boldsymbol{\phi}, \boldsymbol{W}, \boldsymbol{C}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{W}\boldsymbol{\phi}, \boldsymbol{C}).$$

Making predictions

When the parameters of the model are known, the unknown target variable x^* of a subject with image $\boldsymbol{t}^*$ and covariates $\boldsymbol{\phi}_{\setminus x}^*$ can be inferred by inverting the model using Bayes' rule. For a binary target variable $x^* \in \{0, 1\}$ where the two outcomes have equal prior probability, the target posterior distribution takes the form of a logistic regression classifier:

$$p(x^* = 1|\boldsymbol{t}^*, \boldsymbol{\phi}_{\setminus x}^*, \boldsymbol{W}, \boldsymbol{C}) = \sigma\big(\boldsymbol{w}_D^T \boldsymbol{t}^* + w_o\big),$$

where

$$\boldsymbol{w}_D = \boldsymbol{C}^{-1}\boldsymbol{w}_x$$

are a set *discriminative* spatial weights, $\sigma(\cdot)$ denotes the logistic function, and $w_o = -\boldsymbol{w}_D^T(\boldsymbol{W}_{\setminus x}\boldsymbol{\phi}_{\setminus x}^* + \boldsymbol{w}_x/2)$. The prediction of x^* is therefore 1 if $\boldsymbol{w}_D^T\boldsymbol{t}^* + w_o > 0$, and 0 otherwise.

For a continuous target variable with Gaussian prior distribution $p(x^*) = \mathcal{N}(x^*|0, \sigma^2)$, the posterior distribution is also Gaussian with mean

$$\sigma_x^2(\boldsymbol{w}_D^T\boldsymbol{t}^* + b_0), \tag{2}$$

where $b_0 = -\boldsymbol{w}_D^T\boldsymbol{W}_{\setminus x}\boldsymbol{\phi}_{\setminus x}^*$ and $\sigma_x^2 = \big(\sigma^{-2} + \boldsymbol{w}_x^T\boldsymbol{C}^{-1}\boldsymbol{w}_x\big)^{-1}$. The predicted value of x^* is therefore given by (2), which again involves taking the inner product of the discriminative weights $\boldsymbol{w}_D$ with $\boldsymbol{t}^*$.

[5] For notational convenience, we include 1 as a dummy "covariate".

Model training

In practice the model parameters $\boldsymbol{W}$ and $\boldsymbol{C}$ need to be estimated from training data. Given N training pairs $\{\boldsymbol{t}_n, \boldsymbol{\phi}_n\}_{n=1}^N$, their maximum likelihood (ML) estimate is obtained by maximizing the marginal likelihood

$$p\left(\{\boldsymbol{t}_n\}_{n=1}^N | \{\boldsymbol{\phi}_n\}_{n=1}^N, \boldsymbol{W}, \boldsymbol{C}\right) = \prod_{n=1}^N \mathcal{N}\left(\boldsymbol{t}_n | \ \boldsymbol{W}\boldsymbol{\phi}_n, \boldsymbol{C}\right) \tag{3}$$

with respect to these parameters. For the spatial maps $\boldsymbol{W}$, the solution is given in closed form:

$$\boldsymbol{W} = \left(\sum_{n=1}^N \boldsymbol{t}_n \boldsymbol{\phi}_n^T\right)\left(\sum_{n=1}^N \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T\right)^{-1}. \tag{4}$$

Obtaining the noise covariance matrix $\boldsymbol{C}$ directly by ML estimation is problematic, however: For images with J voxels, $\boldsymbol{C}$ has $J(J+1)/2$ free parameters – orders of magnitude more than there are training samples. To circumvent this problem, we impose a specific structure on $\boldsymbol{C}$ by using a latent variable model known as factor analysis [10]. In particular, we model the noise as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{z} + \boldsymbol{\epsilon},$$

where $\boldsymbol{z}$ is a small set of K unknown latent variables distributed as $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \mathbb{I}_K)$, $\boldsymbol{V}$ contains K corresponding, unknown spatial weight maps, and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian distributed error with unknown diagonal covariance $\boldsymbol{\Delta}$. Marginalizing over $\boldsymbol{z}$ yields a zero-mean Gaussian noise model with covariance matrix

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{\Delta},$$

which is now controlled by a reduced set of parameters $\boldsymbol{V}$ and $\boldsymbol{\Delta}$. The number of columns in $\boldsymbol{V}$ (i.e., the number of latent variables K) is a hyperparameter in the model that needs to be tuned experimentally.

Plugging in the ML estimate of $\boldsymbol{W}$ given by (4), the parameters $\boldsymbol{V}$ and $\boldsymbol{\Delta}$ maximizing the marginal likelihood (3) can be estimated using an Expectation-Maximization (EM) algorithm [30]. Applied to our setting, this yields an iterative algorithm that repeatedly evaluates the posterior distribution over the latent variables:

$$p(\boldsymbol{z}_n|\boldsymbol{t}_n, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{\Delta}) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}_n = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)$ and $\boldsymbol{\Sigma} = (\mathbb{I}_K + \boldsymbol{V}^T\boldsymbol{\Delta}^{-1}\boldsymbol{V})^{-1}$, and subsequently updates the parameters:

$$\boldsymbol{V} \leftarrow \left(\sum_{n=1}^N (\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)\boldsymbol{\mu}_n^T\right)\left(\sum_{n=1}^N \left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + \boldsymbol{\Sigma}\right)\right)^{-1}$$

$$\boldsymbol{\Delta} \leftarrow \operatorname{diag}\left(\frac{1}{N}\sum_{n=1}^N (\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)^T - \boldsymbol{V}\frac{1}{N}\sum_{n=1}^N \boldsymbol{\mu}_n(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)^T\right).$$

3 Experiments

In our implementation, we initialize the EM algorithm by using a matrix with standard Gaussian random entries for $\boldsymbol{V}$, and a diagonal matrix with the sample variance in each voxel across the training set for $\boldsymbol{\Delta}$. For continuous target variables, we de-mean the target and use the sample variance as the prior variance σ^2. Convergence is detected when the relative change in the log marginal likelihood is smaller than 10^{-5}.

The method has a single hyperparameter, the number of latent variables K, that we set empirically using cross-validation on a validation set, by optimizing the mean absolute error (MAE) for regression and the accuracy for classification. Running times vary with the size of the training set N, which also influences the selected value of K – in our implementation, typical training runs in the full-brain experiments described below took between 2.8 and 16.3 minutes for $N = 200$ and $N = 1000$, respectively (CPU time for a single selected value of K; Matlab on a state-of-the-art desktop). Once the model is trained, testing is fast: typically 0.01 seconds per subject when trained on $N = 1000$.

Comparing performance of an image-based prediction method with state-of-the-art benchmark methods is hampered by the dearth of publicly available software implementations, and the strong dependency of attainable performance on the datasets that are used [12]. Within these constraints, we conducted the following comparisons of the proposed linear generative method:

Nonlinear discriminative benchmark: As the main benchmark method, we selected the convolutional neural network SFCN proposed in [28], which is, to the best of our knowledge, currently the best performing image-based prediction method. The paper reports performance for age and gender prediction over a wide range of training sizes in preprocessed UK Biobank data (14,503 healthy subjects, aged 44-80 years), using a validation set of 518 subjects and a test set of 1036 subjects. For a training size of 12,949 subjects, the authors report a training time of 65 hours on two NVIDIA P100 GPUs [28]. Although the method uses affinely registered T1-weighted scans as input ("affine T1s"), these are in fact skull-stripped and subsequently bias-field-corrected based on deformable registrations that are also available [4]. Because of this reason, and because the authors report only very minor improvements of their method when deformable T1s are used instead ($\sim 2.5\%$ decrease in MAE for age prediction on 2590 training subjects), we compared our method using both affine and deformable T1s, based on a set-up that closely resembles theirs (validation set of 500 subjects, test set of 1000 subjects).

Linear discriminative benchmark: In order to compare against a state-of-the-art *linear* discriminative method, we selected the RVoxM method [31] because its training code is readily available [1] and its performance is comparable to the best linear discriminative method tested in [28]. RVoxM regularizes its linear discriminant surface by encouraging spatial smoothness and sparsity of its weight maps, using a regularization strength that is the one

hyperparameter of the method. In our experiments, we selected the optimal value of this hyperparameter in the same way as we do it for the proposed method, i.e., by cross-validation on our 500-subject validation set. Typical training times were between 66 and 122 minutes for $N = 200$ and $N = 1000$, respectively (CPU time for a single selected value of the model's hyperparameter; Matlab on a state-of-the-art desktop).

Nonlinear generative benchmark: As a final benchmark, we compared against a variational auto-encoder (VAE) [40] that was recently proposed for age prediction, and that has training code publicly available [2]. It is based on a generative model that is similar to ours, except that its latent variables are expanded ("decoded") *nonlinearly* using a deep neural network, which makes the EM training algorithm more involved compared to our closed-form expressions [26]. In [40], the authors use T1 volumes that are cropped around the ventricular area (cf. Fig. 1 right), and they train their method on ∼200 subjects. We closely follow their example and train both the VAE and the proposed method on similarly sized training sets of warped T1 scans from the UK Biobank, cropped in the same way. There are two hyperparameters in the VAE model (dropout factor and L2 regularization), which we optimized on our validation set of 500 subjects using grid search. The training time for this method was on average 9.40 minutes for N=200 with the optimal set of hyperparameters, using a NVIDIA GeForce RTX 2080 Ti GPU.

For each training size tested, we trained each method three times, using randomly sampled training sets, and report the average test MAE and accuracy results. For gender classification, we used age as a known covariate in $\boldsymbol{\phi}_{\setminus x}$, while for age prediction no other variables were employed. All our experiments were performed on downsampled (to 2mm isotropic) data, with the exception of RVoxM where 3mm was used due to time constraints – we verified experimentally that results for RVoxM nor the proposed method would have changed significantly had the downsampling factor been changed (max difference of 0.32% in MAE between 2mm and 3mm across multiple training sizes between 100 and 1000). Since training code for SFCN is not publicly available, we report the results as they appear in [28], noting that the method was tuned on a 518-subject validation set as described in the paper.

4 Results

Fig. 1 shows examples of the *generative* spatial map $\boldsymbol{w}_x$ estimated by the proposed method, along with the the corresponding *discriminative* map $\boldsymbol{w}_D$. The generative map shows the direct effect age has on image intensities, and reflects the typical age-related gray matter atrophy patterns reported in previous studies [19]. The discriminative map, which highlights voxels that are employed for prediction, is notably different from the generative map and heavily engages white matter areas instead. This illustrates the interpretation problem in discriminative models: the discriminative weight map does not directly relate to changes in neuroanatomy, but rather summarizes the net effect of decomposing

the signal as a sum of age-related changes and a typical noise pattern seen in the training data (1), resulting in a non-intuitive spatial pattern [22].

Fig. 2 shows the performances obtained by the proposed method, compared to the discriminative benchmarks RVoxM and SFCN, for age and gender prediction. Both our method and RVoxM achieve clearly worse results when they are applied to affine T1s compared to deformable T1s, whereas SFCN's performance is virtually unaffected by the type of input data (at least for age prediction with 2590 training subjects – the only available data point for SFCN with deformable T1s [28]). These results are perhaps not surprising, since both our method and RVoxM are *linear* predictors that do not have the same capacity as neural networks to "model away" nonlinear deformations that have not been removed from the input images (even though these are actually known and were used for generating the affine T1s).

Comparing the performances of the different methods, our generative model generally outperforms the linear discriminative RVoxM for both age and gender prediction, except when using very large training sets of affine T1s. For *nonlinear* discriminative SFCN, the situation is more nuanced: For age prediction, SFCN starts outperforming our method for training sets larger than 2600 subjects, while for more moderate training sizes our method achieves better performances when deformable T1s are used. For gender prediction, our method based on deformable T1s is competitive with SFCN even on the biggest training set sizes, although it should be noted that SFCN's results are based on affine T1s as its performance on deformable T1s for gender prediction was not tested[6] in [28].

Finally, Fig. 3 compares the age prediction results of our linear generative model with the nonlinear generative VAE, both trained on cropped deformable T1s. Our method clearly outperforms the VAE for all the considered training sizes, suggesting that, at least when only a few hundred training subjects are available, adding nonlinearities in the model is not beneficial.

5 Discussion

In this paper, we have introduced a lightweight method for image-based prediction that is based on a linear generative model. The method aims to be easier to use, faster to train and less opaque than state-of-the-art nonlinear and/or discriminative methods. Based on our experiments in predicting age and gender from brain MRI scans, the method seems to attain these goals without sacrificing prediction accuracy, especially in the limited training size scenarios that are characteristic of neuroimaging applications.

Although the method presented here is linear in both its causal forward model and in its noise model, it would be straightforward to introduce nonlinearities in the forward model while still maintaining numerical invertibility. This may be beneficial in e.g., age prediction in datasets with a much wider age range than the UK Biobank data used here. The method can also be generalized to longitudinal

[6] Nevertheless, SFCN's gender prediction, based on affine T1s, is reported by its authors to be the best in the literature.

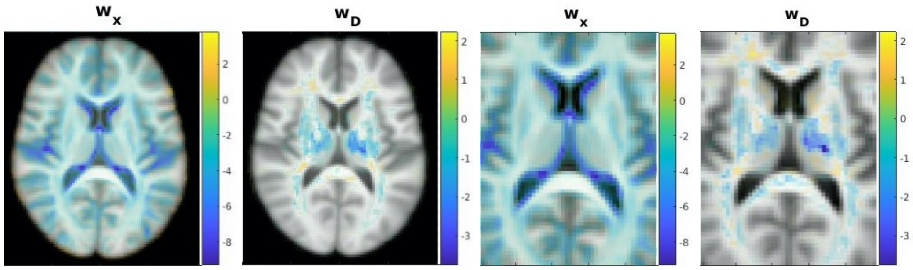

Fig. 1: Examples of *generative* maps $\boldsymbol{w}_x$ encoding age effects vs. the corresponding *discriminative* maps $\boldsymbol{w}_D$ predicting age, obtained on deformable T1s from 300 subjects and overlaid on the average T1 volume. Voxels with zero weight are transparent. Left: results on whole T1 images (used for comparing the proposed method with SFCN and RVoxM). Right: results on cropped T1s (used for comparing with VAE).

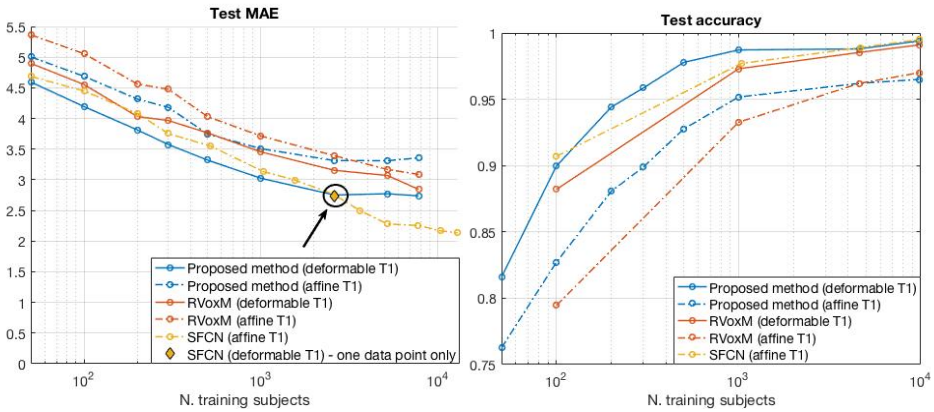

Fig. 2: Comparison of the proposed method, RVoxM and SFCN on an age prediction task (left) and on a gender classification task (right). For each method, results are shown for both affine and deformable T1 input data – except for SFCN for which the result for deformable T1s is only known for age prediction, in a single point (indicated by an arrow at 2590 subjects).

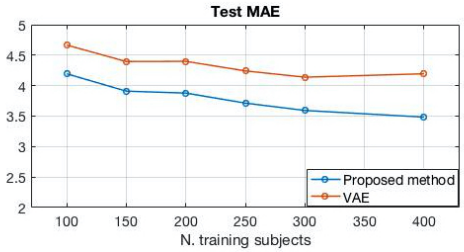

Fig. 3: Test MAE for age prediction obtained by the proposed method and VAE on cropped, deformable T1s.

data, where addressing the intersubject variability in both the timing and the number of follow-up scans is well suited for generative models such as the one proposed here.

Acknowledgments This research has been conducted using the UK Biobank Resource under Application Number 65657. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreements No. 765148 and No. 731827, as well as from the National Institutes Of Health under project numbers R01NS112161 and 1RF1MH117428.

References

1. https://sabuncu.engineering.cornell.edu/software-projects/relevance-voxel-machine-rvoxm-code-release/
2. https://github.com/QingyuZhao/VAE-for-Regression
3. Adebayo, J., et al.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
4. Alfaro-Almagro, F., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage **166**, 400–424 (2018)
5. Arbabshirani, M.R., et al.: Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage **145**, 137–165 (2017)
6. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021)
7. Ashburner, J., et al.: Voxel-based morphometry–the methods. Neuroimage **11**(6), 805–821 (2000)
8. Baehrens, D., et al.: How to explain individual classification decisions. The Journal of Machine Learning Research **11**, 1803–1831 (2010)
9. Batmanghelich, N.K., et al.: Generative-discriminative basis learning for medical imaging. IEEE transactions on medical imaging **31**(1), 51–69 (2011)
10. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4, chap. 12. Springer (2006)
11. Chung, M., et al.: A unified statistical approach to deformation-based morphometry. NeuroImage **14**(3), 595–606 (2001)
12. Cole, J.H., et al.: Quantification of the biological age of the brain using neuroimaging. In: Biomarkers of human aging, pp. 293–328. Springer (2019)
13. Davatzikos, C., et al.: Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage **14**(6), 1361–1369 (2001)
14. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6), 659–667 (2014)
15. Domingos, P., et al.: On the optimality of the simple bayesian classifier under zero-one loss. Machine learning **29**(2), 103–130 (1997)
16. Ellis, K.A., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
17. Erhan, D., et al.: Visualizing higher-layer features of a deep network. University of Montreal **1341**(3), 1 (2009)
18. Fischl, B., et al.: Measuring the thickness of the human cerebral cortex from magnetic resonance images. PNAS **97**(20), 11050 (2000)
19. Fjell, A.M., et al.: High Consistency of Regional Cortical Thinning in Aging across Multiple Samples. Cerebral Cortex **19**(9), 2001–2012 (2009). https://doi.org/10.1093/cercor/bhn232
20. Ghassemi, M., et al.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)
21. Glasser, M.F., et al.: The human connectome project's neuroimaging approach. Nature neuroscience **19**(9), 1175–1187 (2016)

22. Haufe, S., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage **87**, 96–110 (2014)
23. He, T., et al.: Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage **206**, 116276 (2020)
24. Jack Jr, C.R., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
25. Kaufmann, T., et al.: Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nature neuroscience **22**(10), 1617–1623 (2019)
26. Kingma, D.P., et al.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
27. Ng, A.Y., et al.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in neural information processing systems. pp. 841–848 (2002)
28. Peng, H., et al.: Accurate brain age prediction with lightweight deep neural networks. Medical image analysis **68**, 101871 (2021)
29. Ras, G., et al.: Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research **73**, 329–397 (2022)
30. Rubin, D.B., et al.: Em algorithms for ml factor analysis. Psychometrika **47**(1), 69–76 (1982)
31. Sabuncu, M.R., et al.: The Relevance Voxel Machine (RVoxM): A Self-Tuning Bayesian Model for Informative Image-based Prediction. IEEE transactions on medical imaging **31**(12), 2290–2306 (2012)
32. Schulz, M.A., et al.: Deep learning for brains?: Different linear and nonlinear scaling in uk biobank brain images vs. machine-learning datasets. BioRxiv p. 757054 (2019)
33. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Shrikumar, A., et al.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
35. Simonyan, K., et al.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations (2014)
36. Smilkov, D., et al.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
37. Springenberg, J.T., et al.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
38. Sundararajan, M., et al.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
39. Varol, E., et al.: Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–548. Springer (2018)
40. Zhao, Q., et al.: Variational autoencoder for regression: Application to brain aging analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 823–831. Springer (2019)